University of Szeged
Doctoral School in Linguistics
PhD Program in Theoretical Linguistics

Danang Satria Nugraha

A Theoretical and Corpus Linguistics Study of the Light Verb Constructions: Empirical Data from Indonesian

Summary of Doctoral Dissertation

Supervisor:
Veronika Vincze, PhD

Szeged, 2025

Abstract: This dissertation investigates Indonesian Light Verb Constructions (LVCs) through the integrated lenses of theoretical linguistics and corpus-based analysis. The primary part of the study provides an empirical foundation, drawing on four corpora to assess the frequency and distribution of LVCs. K-means clustering reveals three natural groupings (low-, medium-, and high-frequency), with Hypothesis 1 supported by significant cross-corpus rank consistency (Spearman's $r_s = 0.891, p <.001$), validating the temporal and genre-stable nature of LVC frequency patterns. Hypothesis 2 confirms a significant deviation from Zipfian expectations, aligning more closely with Zipf–Mandelbrot law. Additional modeling supports the Menzerath–Altmann law (morpheme-based fit), while vocabulary dynamics are elaborated through Heaps' Law and Baayen's productivity metrics. Lexical drift is capture diachronically via Yule's K and KL Divergence, while entropy measures underscore shifting lexical concentration. A contrastive analysis with Altmann's (1967) seminal dictionary-based lexical model of Indonesian reveals substantial structural divergence in morpheme density, type distribution, and correlation behavior. Hypothesis 3 is confirmed through morpho-semantic-syntactic stratification across clusters, identifying a structurally asymmetric, gradiently layered LVC system. The secondary part classifies verb elements into True Light Verbs and Vague Action Verbs using aktionsart diagnostics. Machine learning (Naïve Bayes and Random Forest) highlights frequency and semantic parameters as strong predictors of verb productivity. The final part analyzes noun components, distinguishing stative and eventive types based on temporal features. Findings indicate that stative interpretations are largely noun-driven, while eventive reading emerge from verb-noun interaction and distributional patterns. This research offers a theoretically informed, data-driven typology of Indonesian LVCs, contributing to corpus linguistics and the broader modeling of LVC systems in underdescribed languages.

1. Introduction

An exploration to Light Verb Constructions (henceforth: LVCs) in Indonesian is somehow limited quantitatively and qualitatively. While the phenomenon undoubtedly merits in-depth investigation, the available studies have thus far provided a limited scope of inquiry. This limitation is further compounded by the status of Indonesian as one of the world's many low-resource languages (henceforth: LRLs). As noted by Singh (2008), Cieri *et al.* (2016), and Tszetkov (2017), LRLs are often characterized by being less studied, resource-scare, less computerized, less privileged, less commonly taught, or of low density. In the context of Natural Language Processing (hereafter: NLP), a field that experienced a major shift from rule-based to statistical-based techniques in the 1990s, LRLs like Indonesian face challenges as most of today's NLP research focuses on a small fraction (around 20) of the world's 7,000 languages, making it difficult to directly apply statistical methods due to data scarcity (Maxwelll-Smith *et al.*, 2022; Magueresse *et al.*, 2020; Sebastian *et al.*, 2022). This scarcity of resources, including annotated corpora and NLP tools, necessitates a research design that incorporates both corpus-based analysis and introspective linguistic insights from native speakers. Such a combined approach aims to maximize the information available for the study of LVCs in Indonesian, acknowledging the limitations inherent in LRL research. Consequently, a comprehensive understanding of complex linguistic phenomena such us LVCs within the Indonesian linguistic system remains elusive, underscoring the need for further exploration and systematic analysis to elucidate the full range of their morpho-syntactic and morpho-semantic properties.

LVCs, by their nature, represent a linguistically intriguing phenomenon prevalent across diverse range of languages (*see* Example 1.1 and 1.2). Contrary to the conventional expectation that such constructions are confined to inflectional languages (Butt, 2010), LVCs have also been recognized in other type of languages, demonstrating their remarkable adaptability to varying morphological systems (e.g., Bygi *et al.*, 2018; Nagy T. *et al.*, 2020; Miyamoto & Kishimoto, 2016; Nugraha & Vincze, 2024; Xu *et al.*, 2022). This versatility equips LVCs with the capacity to conform to the grammatical structures of diverse linguistic systems while simultaneously providing speakers with a versatile tool for expressing a wide spectrum of meanings. In major cases, the nucleus of an LVC comprises a semantically bleached verb, often conveying a general meaning akin to ‘do’ or ‘make’ in English, in conjunction with another element, typically a noun, which conveys the specific meaning of the construction (cf., Giparaitė, 2015, 2024; Ronan, 2014, 2019; Ronan & Schneider, 2015).

(1.1) Indonesian (Nugraha, 2024: 104)
memberikan *kuliah*
give-TR.PRED lecture-OBJ.P
‘give a lecture’

(1.2) Hungarian (Vincze, 2011: 283)
kutatást *folytat*
research-ACC do
‘do research’

According to Mel’čuk’s (2022) formal definition, the LVCs is consisting of a support verb ($V_{(support)}$) and a noun with a semantic predicate ‘σ’, where the resulting VN phrase mirrors the meaning of N‘σ’ alone. In (1.1), Indonesian ‘*memberikan kuliah*’ (to give a lecture), the verb ‘*memberikan*’ (to give) functions as the $V_{(support)}$, providing the necessary grammatical scaffolding for the noun ‘*kuliah*’ (lecture), which carries the core semantic predicate of ‘lecturing’. The resulting phrase, ‘*memberikan kuliah*’, effectively conveys the meaning ‘to lecture’, equivalent to the semantic content embedded within ‘*kuliah*’ itself. This construction exemplifies how Indonesian employ light verb to create LVC, where the light verb contributes minimal semantic content beyond grammatical function, allowing the noun to take center stage. Thereafter, in Hungarian (1.2) ‘*kutatást folytat*’ (do research), ‘*folytat*’ (to do/carry out) serves as the $V_{(support)}$, while ‘*kutatást*’ (research), in the accusative case, provides the semantic predicate. The phrase ‘*kutatást folytat*’ translates to ‘do research’, effectively verbalizing the concept encapsulated within the noun ‘*kutatást*’. This construction demonstrates how Hungarian utilizes LVCs to express activities or processes associated with the noun, highlighting the flexibility of this structure in capturing a wide range of semantic nuances.

Considering some of this background, especially the universal nature of LVCs in various languages, to support the aim of exploring LVCs in Indonesian, the corpus linguistics approach employed in the current study. It underscores the significance of empirical analysis in understanding those language phenomena in the language-specific patterns. By examining a large dataset of Indonesian LVCs, we were able to identify patterns and trends that would have been difficult to discern through introspection or intuition alone. This methodology provides a solid foundation for understanding the complex interplay between the various light verb and their accompanying nouns, revealing the multifaceted nature of these constructions and their role in conveying meaning within the Indonesian grammatical context.

2. Method of the study

2.1 Research design and approach

The present study delved into the landscape of LVCs in Indonesian by employing a research design that integrated a corpus-based approach with principles from theoretical linguistics (see Figure 1). This integrated approach facilitated a comprehensive examination of LVCs, encompassing their frequency, distribution, underlying mechanisms, and linguistic contribution within the language (Biber, 2012; Egbert *et al.*, 2020). A large and representative corpus of Indonesian text constituted the core of the research design. This corpus provided the empirical foundation for the investigation, yielded a substantial amount of authentic language data in which LVCs could be observed in their natural context. Corpus analysis tools and techniques were employed to identify and categorize LVC types within the corpus. This process involved quantifying LVC occurrences and analyzing their distribution across different grammatical conditions. Additionally, the analysis also examined the inherent features of LVCs through their verb and noun elements.

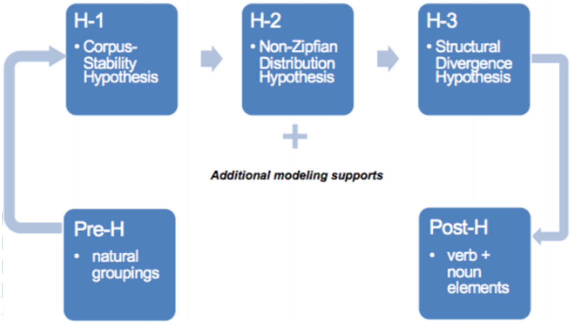

H-1
• Corpus-Stability Hypothesis
H-2
• Non-Zipfian Distribution Hypothesis
H-3
• Structural Divergence Hypothesis
+
Additional modeling supports
Pre-H
• natural groupings
Post-H
• verb + noun elements

Figure 1. Research design in the light of hypotheses arrangement.

2.2 Materials

First, data. The data for this study comprises 942 types of Indonesian LVCs. Specific instantiations of these LVCs can be observed in examples (1.1) and (1.2). These instances, along with their frequency and underlying semantic and syntactic features, constitute the empirical data that form the research object of the current study. In particular, the frequency data for each LVCs, within its initial categorization as *hypothetical* or *genuine*, obtained from four corpora in this study, serves as a crucial foundation for the quantitative analysis.

Second, data sources. The selection of data sources for this research was predicated upon several critical criteria. These criteria encompassed: (i) the sources' established reputation for credibility and rigorous data collection methodologies, (ii) the direct relevance of their data to the research topic under investigation, and (iii) the comprehensive coverage they provided of the subject area. The selected sources underwent a rigorous evaluation process, focusing on data quality, accessibility, and alignment with the project timeline. Following careful deliberation, ILCC, SLIC, IDC, and IWC were chosen as the data sources, as they demonstrated optimal performance across these evaluative parameters. As a further explanation, statistical detail on these four corpora can be found in Table 1.

Table 1. Details identification of the selected corpus.

No	Aspect	ILCC	SLIC	IDC	IWC
1.	Description	The Indonesian – Leipzig Corpora Collection is a mixed corpus based on material 2013. The majority of text comes from internet websites.	This monolingual corpus consists of Indonesian texts retrieved from a variety of internet sources.	The Indonesian Web Corpus (idTenTen) is an Indonesian corpus made up of texts collected from the internet.	The Indonesian web corpus (idWaC) is an Indonesian corpus made up of texts collected from the Internet.
2.	Name	Ind_mixed_2013	SEAlang Library Indonesian Text Corpus	idTenTen	idWaC
3.	Language	Indonesian	Indonesian	Indonesian	Indonesian
4.	Genre	Text-based (Online)	Text-based (Online)	Text-based (Online)	Text-based (Online)
5.	Year	2013	2010	2020	2012
6.	Sentences	74,329,815	2,242,565	258,435,545	6,003,769
7.	Types	7,964,109	5,000,000	3,678,192,045	90,120,046
8.	Tokens	1,206,281,985	15,763,657	4,432,864,160	109,236,814
9.	Token/sent.	16.23	7.03	17.15	18.19
10.	Link to corpus	https://corpora.uni-leipzig.de/?corpusId=ind_mixed_2013	http://sealang.net/indonesia/corpus.htm	https://www.sketchengine.eu/idtenten-indonesian-corpus/	https://www.sketchengine.eu/indonesianwac-corpus/

Third, unit of analysis. For the *pre-h, H-1, H-2,* and *H-3* arrangements, which explores LVC frequency and distribution, the unit of analysis are individual LVC occurrences within the all corpora. A sample of the unit of analysis for these explorations is presented in Table 2.

Table 2. Sample of data-unit for frequency analysis..

No.	LVC	Total (Σ-Frequency)	Total (Z-Score)	PMW	Rank	Cluster
1.	*memberikan kuliah* 'to give a lecture'	12,532	-0.195	628.989	308	1
2.	*mengambil kesempatan* 'take a chance'	24,931	0.086	1252.203	206	2
3.	*membuat keputusan* 'make a decision'	114,153	2.106	5733.532	40	3

The *post-H* analyses are concerned with an investigation of the linguistic nature of LVCs. Here, the unit of analysis is broadened to consider not just the LVC as a composite structure, but also its inherent component parts, namely verb and noun elements. This refined approach enables a detailed exploration of how aktionsart affects the characteristics inherent in Indonesian LVCs. By adopting this expanded unit of analysis, which foregrounds the pivotal role of the verb and noun, the study presents an extended examination of their internal structure. Consider the Example in (1.3) to (1.5) as follows.

(1.3)	*Terutama*	*dipakai*	*oleh*	*kita*	*dalam*	***mengambil***	***keputusan***	*praktis.*
	especially-ADV	use-PASS	by-ADP	us-3PL	in	take-TR.	decision-OBJ.P	practical

'We primarily use this in making practical decision.'

(1.4)	*Sebab*	*pihaknya*	*masih*	*harus*	***melakukan***	***kajian***	*dan*	*analisa*	*kasus.*
	because	their party	still	have	do-TR.	study-OBJ.P	and	analysis	case

'That's because they still need to conduct a study and analysis of the case.'

(1.5)	*Umar*	***memberikan***	***saran***	*kepada*	*Hz.*
	Umar-PROPN.A.SBJ.TOP	give-TR	advice-OBJ.P	to-ADP	Hz-PROPN.BEN.SG

'Umar gave advice to Hz.'

2.3 Procedures

First, data collecting procedures. Data collection procedures involve (a) query search assemblage, (b) inclusion/exclusion criteria, and (c) frequency extraction. The initial stage includes preparing two matrices: a hypothetical matrix, which comprises light verbs from prior studies on various languages, and a genuine matrix derived from introspective input from native Indonesian speakers. Both matrices underwent validation within four corpora, requiring a minimum of one occurrence each (Baker & Egbert, 2016). Consequently, the genuine dataset was reduced from 102 to 101 items, while the hypothetical dataset decreased from 900 to 841 items, resulting in a total of 942 validated types of light verb constructions (LVCs) for frequency extraction. The analysis, inspired by McEnery and Hardie (2012), provided raw counts of LVC occurrences. To enable meaningful comparisons between corpora of varying sizes, these raw frequencies were normalized. As Baker *et al.* (2006: 75) note, this can be done using percentages or occurrences per-million-words (PMW), helping to standardize comparisons and mitigate the impact of corpus size on frequency counts.

Second, data analysis procedures. The procedures comprise (a) frequency analysis, (b) distribution analysis, and (c) verb and noun elements analyses. To analyze the natural circulation of LVCs, a process of frequency clusterization was employed by utilizing K-means Clustering performed in R-Studio application. Furthermore, frequency data based on natural clusters is then analyzed for its distribution. In light of *H-1* about the corpus stability, Spearman's rank correlation to discern a pattern in the distribution of LVCs across the four corpora. According to *H-2*, the Zipfian distribution analysis is performed to assess whether the frequency distribution follows or deviates from Zipf's Law. The analysis of distribution also focuses on six key empirical laws: (a) the Menzerath-Altmann Law, positing an inverse relationship between the size of a linguistic construct and the size of its constituent; (b) Good-Turing Frequency Estimation, a method for estimating the probability of unseen or low-frequency events; (c) Yule's K, a measure of lexical diversity and concentration; (d) Heaps' Law, which describes the growth of vocabulary size in relation to text length; (e) Baayen's morphological productivity framework, evaluating tendencies of the morphological productivity within LVCs; and (f) Shannon Entropy, quantifying lexical unpredictability and distributional evenness within LVCs dataset.

Moreover, complementary instruments were developed to investigate the distribution of Indonesian LVCs across six predefined contexts in the light of *H-3*. The first context, measured by the M-1 parameter, traced LVC distribution based on nominal morphological features in the noun element (Mel'čuk, 2006; Blevins, 2016; Lieber & Štekauer, 2011). The second context, evaluated through the M-2 parameter, examined LVC distribution based on the noun element's primary feature, specifically the conceptual skeleton of SUBSTANCE (Lieber, 2004, 2010; Dal & Namer, 2015). The third context utilized the S-1 parameter to analyze LVC distribution concerning synonymity (Cruse, 1986; Andreou, 2017; Hrenek, 2021). The fourth context, assessed with the S-2 parameter, focused on prototypicality (Coleman & Kay, 1981; Singleton, 2000; Aikhenvald, 2006, 2007, 2017). The last two contexts included Sx-1 for transitivity (Hopper & Thompson, 1980; Malchukov, 2006; Kittilä, 2006) and Sx-2 for valency frames (Kettnerová, 2023; Vincze, 2014). These analyses describe the distribution tendency of LVCs in Indonesian grammatical contexts. Additionally, aktionsart and temporal properties of verb and noun elements in Indonesian LVCs were examined using specific metrics, relying significantly on the author's intuition as a native speaker. The matrices were derived from Binnick's (1991) work for verb elements and Pompei *et al.*'s (2023) study for noun elements.

3. Results

3.1 Frequency of Indonesian LVCs

First, frequency groups from distinct natural clusters. The current K-means clustering analysis in R-Studio was conducted using the following parameters: the optimal number of clusters (k) was determined using the three validation measures; the initialization method was employed for calculating distances between data points, as no compelling reasons were found to necessitate alternative distance measures; and the iteration settings, including the maximum number of iterations and convergence criteria, were adjusted to optimize the clustering process and ensure the stability of the resulting clusters. In detail, Figure 2 presents the cluster plot resulting from a K-means clustering analysis. In the context of analyzing LVCs, this visualization provides empirical insights into the central tendencies or representative characteristics of each identified cluster. It outlines three distinct clusters (marked as Q1, Q2, and Q3), each associated with a unique centroid as presented in Table 3. These centroids serve as reference points in a multi-dimensional space defined by the variables of LVCs frequency within ILCC, SLIC, IDC, and IWC. Their numerical values in each centroid offer insights into the distinguishing characteristics of each cluster.

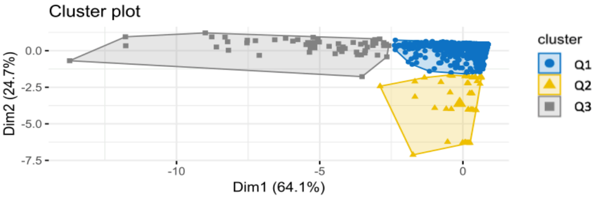


Figure 2. Cluster plot of the results of K-means clustering.

In Cluster Q1, the cluster is notable for its low negative values across all variable, i.e., ILCC (-0.19), SLIC (-0.21), IDC (-0.20), and IWC (-0.19). It suggests a grouping of LVCs sharing a pronounced tendency towards the linguistic properties associated with these variables. Based on these centroid coefficients, Q1 covers all low-frequency LVCs. In contrast, Cluster Q3 is characterized by a strikingly high value for all variable, i.e., ILCC (2.92), SLIC (0.15), IDC (3.20), and IWC (2.99). Grounded on these centroid coefficients, Q3 coverings entirely high-frequency LVCs. Lastly, cluster Q2, the cluster displays a relatively mixture profile. Three variable values are negative, i.e., ILCC (-0.08), IDC (-0.19), and IWC (-0.09); and one value is positive as in SLIC (3.60). Q2 is the middle cluster in the frequency distribution of LVCs in the four corpora analyzed. Accordingly, the first cluster comprises 841 LVCs, the second encompasses 46 LVCs, and the third contains 55 LVCs.

Table 3. Centroid of clusters.

Cluster	ILCC	SLIC	IDC	IWC
Q1	-0.19	-0.21	-0.20	-0.19
Q2	-0.08	3.60	-0.19	-0.09
Q3	2.92	0.15	3.20	2.99

Second, cross-corpus frequency analysis of linguistically defined LVCs. To rigorously assess Hypothesis 1, which posits statistically significant correlations in the frequency of specific LVC types across the four corpora, this subsection presents a quantitative analysis employing statistical correlation tests. The analysis explores the consistency of LVC type distribution across the ILCC, SLIC, IDC, and IWC corpora.

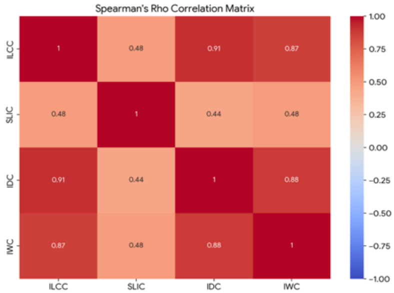


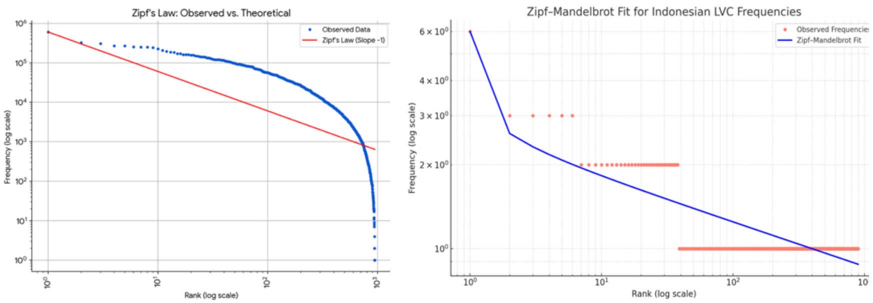
Figure 3. Spearman’s rank correlation to discern pattern in the distribution of LVCs across the four corpora.

As illustrated in Figure 3, the analysis revealed robust positive correlations between several variables, suggesting a degree of systematicity in LVC usage. Specifically, a strong positive association was observed between ILCC and IDC (r_s = .91, p < .001, N = 942), indicating that the relative frequencies of LVCs and other lexical items in these two corpora exhibit a high degree of concordance. This suggests that LVCs, along with other lexical components, tend to occur with similar relative frequencies in both ILCC and IDC. A similarly strong positive correlation was found between ILCC and IWC (r_s = .87, p < .001, N = 942), further supporting the notion that LVC usage patterns are broadly consistent between these corpora. Furthermore, IDC and IWC demonstrated a very strong positive correlation (r_s = .88, p < .001, N = 942), suggesting that these two corpora share a particularly high degree of similarity in their lexical distributions, implying a shared preference or constraint in LVC selection and frequency. These findings, taken together, suggest a substantial degree of covariation in the occurrence of LVCs, across these three corpora, implying that common underlying factors, such as shared linguistic development or similar stylistic preferences, may influence their lexical composition.

Third, Zipfian Distribution in the frequency profiles of LVCs as multiword expressions. The plot reveals a general downward trend, as demonstrated in Figure 4(a), indicating an inverse relationship between the rank and the frequency of LVCs. This aligns with the basic principle of Zipf’s Law[1]: more frequent LVCs tend to have higher ranks, and vice-versa. The observed data points show a *roughly linear* pattern, suggesting that Zipf’s Law captures a significant aspect of the frequency distribution of LVCs. While the overall trend is linear-like, there are noticeable deviations from a perfect straight line, particularly at the lower frequency (high rank) end of the plot. The deviations of the observed data points from the line highlight the ways in which the actual frequency distribution of LVCs differs from the theoretical Zipfian distribution. Moreover, for the high-frequency end (low rank), the data points representing the most frequent LVCs (those with low ranks, towards the left of the plot) tend to align more closely with the theoretical Zipf’s Law line (red line). This suggests

[1] *See* Altmann (1985, 1997, 2025), Köhler *et al*., (2005), Oakes (2019), and (Zipf, 2013).

that Zipf's Law provides a relatively good fit for the most common LVCs. Additionally, for the low-frequency end (high rank), the data points representing the less frequent LVCs (those with high ranks, towards the right of the plot) show a more pronounced deviation from the theoretical line. Specifically, there is a clear mark for increased scatter and slight upward curvature. As the increased scatter, there is more variability in the frequency of low-frequency LVCs compared to high-frequency LVCs. The points are more dispersed and do not adhere as closely to a linear trend. As the slight upward curvature, the observed data points tend to be located *above* the theoretical Zipf's Law line in this region. This indicates that the low-frequency LVCs are somewhat *more frequent* than predicted by a strict Zipfian distribution. In other words, there are more low-frequency LVCs than expected if Zipf's Law held perfectly.



(a) (b)

Figure 4: Result of (a) log-log plot analysis on observed vs. theoretical LVCs (Zipfian distribution) and (b) Zipf-Mandelbrot fit for LVC frequencies.

Furthermore, the Zipfian analysis reveals a notable discrepancy in the *tail region of the distribution,* where the majority of LVC types occur infrequently. This tail-heavy structure—common in high productive linguistic domains—undermines the linearity assumed by the pure Zipfian form. To address this, as illustrated in Figure 4(b), the Zipf–Mandelbrot law[2] was applied as an extension, incorporating an additional shift parameter to account for non-linear head and tail behavior. The model yielded parameters $a \approx 2.59$, $b \approx 0.159$, and $c \approx -0.995$, producing a more accurate fit to the empirical data, particularly in the low-frequency ranks. The low exponent b indicates a relatively flat distribution, consistent with a system where many LVCs share similar, modest usage rates. The slight negative value of c further suggests an adjustment for the *frequency plateau* observed in the lower ranks—deviation that could not be captured by the simples Zipf model. Therefore, the observed deviation in low-frequency LVCs is not merely statistical noise, but rather a manifestation of the *rich tail of the lexicon,* where creative, emergent, or context-specific constructions reside. These constructions are functionally valuable even if infrequent, and their proliferation skews the distribution away from the canonical Zipfian slope. The Zipf–Mandelbrot enables a more faithful modeling of this phenomenon by capturing the *flattened tail* without distorting the overall rank-order hierarchy. This refinement supports a broader view of the lexicon as hierarchically organized, where frequency patterns are shaped by semantic granularity, inherently linguistic context, and morphosyntactic affordances.

[2] *See* Altmann and Gerlach (2016), Manin (2009), Piantadosi (2014), and Popescu *et al*. (2010).

3.2 LVCs' distribution in relation to empirical laws of language

First, internal structure of LVCs and the Menzerath–Altmann Law. The Menzerath–Altmann Law, a cornerstone of quantitative linguistics, describes an inverse relationship between the size of a linguistic construct and the size of its constituent elements. Often summarized as "*the greater whole, the smaller its parts*," this principle has been applied across various levels of linguistic organization, from phonology to syntax. In the present study, the law is applied to Indonesian LVCs, which serve as productive multiword expressions comprising a verbal element and a nominal element. This analysis investigates whether longer LVCs are composed of less frequent constituents—operationalized here through the inverse of observed frequency (`Inverse_TOTAL`). Four models are developed to reflect different conceptions of *length*: word-based, morpheme-based, syllable-based, and phoneme-based. Each model is fitted using Menzerath–Altmann function $y = a\ xb\ e - cx$, and their comparative fits are assessed using the coefficient of determination (R^2). By visualizing and comparing these models within a unified log-scaled framework (*see* Figure 5), this analysis not only tests the applicability of the law in the context of LVCs but also evaluates which linguistic unit best captures the principle of constituent economy in the Indonesian lexicon.

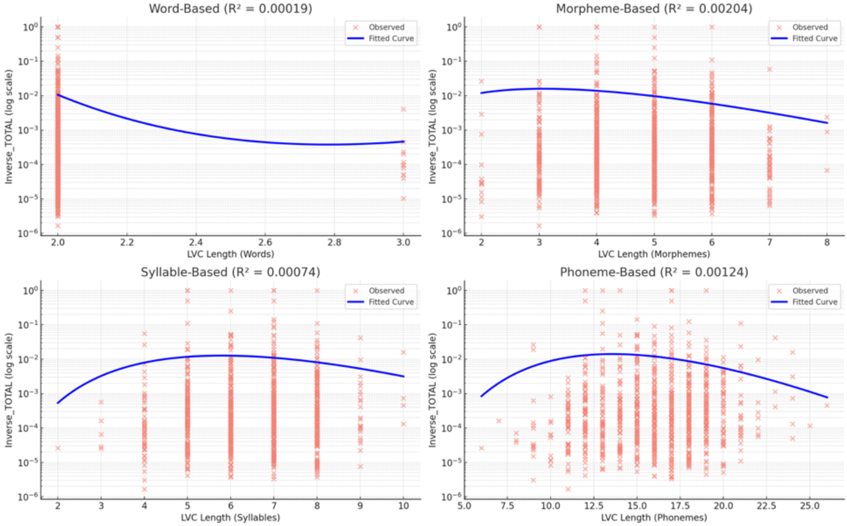


Figure 5. Testing the Menzerath-Altman Law using word–, morpheme–, syllable–, and phoneme–based LVC lengths.

The comparative of the Menzerath–Altmann Law across word-, morpheme-, syllable-, and phoneme-based measures offers nuanced insight into the structural dynamics of Indonesian LVCs. Empirically, the morpheme-based model emerges as the most robust representation, providing the clearest fit to the inverse law with $R^2 \approx 0.00204$. The morpheme-based model follows closely, suggesting that both morpho-semantic and phonological considerations are essential to understanding constituent complexity. The syllable-based model, while partially aligned, underperforms relative to morphemes and phonemes, and the word-based model offers the weakest support. These findings validate the theoretical premise that linguistic economy is best observed at structural—not surface—levels of representation.

Second, an application of Good-Turing Frequency Estimation. In linguistic applications, such as corpus-based analysis of multiword constructions or lexical items, the Good–Turing method serves as a robust tool for estimating the likelihood of unobserved but theoretically plausible expressions. Rather than relying solely on the empirical frequency of each item, the technique models how often items of a given frequency (e.g., those occurring once, twice, etc.) appear and uses that meta-distribution to redistribute probability mass. This particularly important in context involving highly skewed frequency distributions, such as LVCs, where a small number of collocations dominate usage while the long tail consists of numerous infrequent or unattested items. The Good–Turing method not only smooths frequency estimates but also allows for the quantification of unseen types—those constructions which are not present in the corpus but are presumed to exist within the larger linguistic system. Given the productivity and compositionality of LVCs in Indonesian, applying the Good–Turing framework provides a perspective on the structural richness of the lexicon and the completeness of the current dataset.

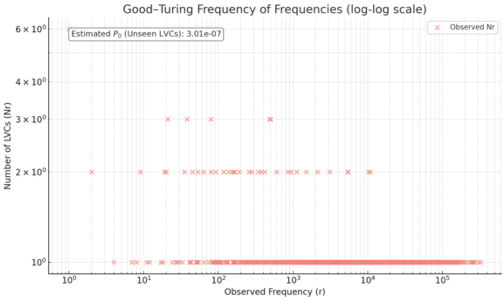


Figure 6. Good-Turing frequency-of-frequency plot in log-log scale, which is visualizes how many LVCs occur with each observed frequency (*r*)

One of the implications of the Good–Turing method is ability to estimate the probability mass associated with unseen events—in this case, LVCs that presumably do not appear in the dataset at all. Based on the presence of six hapax legomena (LVCs observed exactly once), the probability of encountering an unseen LVCs was estimated to be approximately P0 ≈ 3.01×10-7. This is an extremely small probability, suggesting that the current dataset is highly comprehensive and captures nearly the entire productive space of observable LVCs within the sample. Using this probability estimate in conjunction with the number of observed LVC types, the projected number of unseen LVCs in the language was calculated to be approximately 0.00028. This minuscule figure indicates that the likelihood of encountering a novel LVC not present in the dataset is exceedingly low—an important finding for evaluating corpus completeness and model generalizability. From practical standpoint, this result affirms that the observed LVC inventory saturates the data distribution and that further data collection is unlikely to yield a substantially different set of constructions. Theoretically, it supports the notion that the majority of structurally plausible LVCs in the language are already attested, thereby strengthening the readability of distributional analyses based on this corpus. In sum, the application of Good–Turing estimation has provided both a corrective lens of interpreting rare constructions and a quantitative basis for asserting the sufficiency of the dataset in representing the Indonesian LVC system.

Third, notable insights from Yule's K. In this study, Yules' K is employed to trace changes in the distributional density of LVCs across four Indonesian corpora: SLIC (2010), IWC (2012), ILCC (2013), and IDC (2020). This diachronic perspective allows us to observe whether LVC usage becomes increasingly diverse or repetitive over time, revealing patterns in linguistic innovation, formulaicity, and register-driven regularity. Given that Yules' K increases as repetition intensifies, a rising trend in K-values would indicate lexical entrenchment, whereas a decline would reflect diversification and structural expansion of LVCs (*see* Figure 7). The present analysis reveals substantial fluctuation in Yules' K across the decade, suggesting non-linear changes in lexical density. These variations can be tied to shifts in corpus composition, genre distribution, or broader sociolinguistic dynamics such as the digitization of discourse and evolving syntactic preferences. Crucially, this approach treats lexical concentration not as a static feature of the language but as a historically contingent outcome of communicative, stylistic, and structural pressures acting on collocational behavior.

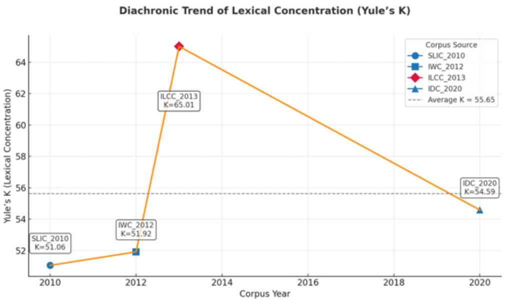


Figure 7. Diachronic trend of lexical concentration based on Yule's K.

The computed Yule's K values offer a quantifiable basis for comparing lexical repetition across the four corpora. The earliest one, SLIC (2010), exhibits the lowest concentration with a Yule's K of 51.06, followed closely by IWC (2012) at 51.92. these values suggest a period of high lexical diversity, characterized by a relatively balance distribution of LVC types and a lower recurrence of fixed expressions. However, this pattern shifts markedly in ILCC (2013), where the Yules' K rises sharply to 65.01—the highest among all four corpora. This peak in lexical concentration suggests a temporary phase of lexical saturation of formulaic reliance, possibly driven by the nature of the ILCC corpus content (e.g., formal documents, educational materials) or a stylistic preference for repeated constructions. In contrast, the IDC (2020) corpus shows a modest reduction in concentration, with a Yule's K of 54.59. This return to lower repetition levels may reflect the influence of more recent language usage norms, including informalization, digital genre discourse. The diachronic curve, as visualized in the enhanced plot, captures a non-monotonic trend: lexical diversity increases from 2010 to 2020, peaks in formulaicity in 2013, and then re-diversifies by 2020.

Fourth, an examination through Heaps' Law. Given the data structure of the present study, an adaptation of Heaps' Law can be employed to investigate the relationship between the number of distinct lexical varieties (LVC types) and their corresponding cumulative frequencies, wherein each row is construed as a discreate type representing a specific lexical variety of LVC and the total column denotes its frequency of occurrence. Consequently, several analytical procedures become feasible: the summation of the total column yields the aggregate number of tokens within the dataset, while the total count of rows directly furnishes the overall number of distinct lexical varieties (LVC types). Although a direct application of Heaps' Law in its conventional cumulative formulation is not strictly feasible, an exploration of the relationship between the rank of a lexical variety (as indicated by the RANK column) and its frequency (TOTAL column) can be undertaken to ascertain the presence of a power-law-like distribution, a phenomenon intrinsically related to Heaps' Law.

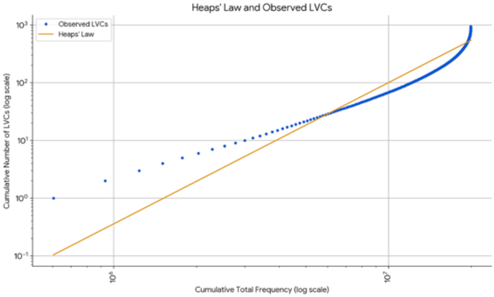


Figure 8. Atypical growth of LVC varieties compared to Heaps' Law.

Figure 8 offers a comparative perspective on the cumulative distribution of observed LVCs against the theoretical trajectory predicted by Heaps' Law. The empirical data, represented by discrete blue markers, illustrates the accumulated number of distinct LVC types as a function of the cumulative total frequency within the dataset. Juxtaposed against this empirical observation is the orange line, delineating the expected growth in lexical variety as posited by Heaps' Law, parameterized by the estimated values of *K* (7.387577325128133e-16) and *β* (2.44776617974269). A visual inspection reveals a marked divergence between the observed LVC accumulation and the anticipated growth curve dictated by the Heapsian model. In typical linguistic corpora, Heaps' Law manifests as a decelerating increase in the number of unique lexical items with increasing corpus size, characterized by an initial rapid expansion followed by a gradual plateauing. However, the depicted empirical data exhibits a distinct pattern, suggesting a fundamentally different relationship between frequency accumulation and the emergence of novel LVC forms within this specific dataset. This initial observation necessitates a more granular examination of the estimated parameters and the underlying characteristics of the data that might account for this deviation from the conventional Heapsian growth pattern.

Fifth, an examination through Baayen's framework. To complement the structural and frequency-based analyses of LVCs, this study incorporates Baayen's morphological productivity framework[3]. This approach enables an evaluation of linguistic creativity and lexical innovation by focusing on the distribution of hapax legomena—types that occur only once in the dataset. Three key metrics are utilized: realized productivity ($P = V/N$), expanding productivity ($S = V_1/V$), and potential productivity ($V = V_1/N$, synonymous with P in this case). The analysis is applied to both LVC clusters and diachronic corpus segments to assess how morphological richness is distributed across linguistic groupings and over time. The use of log-scaled heatmaps enhances the interpretability of these data, as it enables subtle

[3] *See* Baayen (1989, 1992, 2009), Baayen and Lieber (1991).

differences in productivity values—often spanning multiple orders of magnitude—to be effectively visualized and compared. This method is particularly useful in the present study, where most productivity measures are extremely low but still theoretically significant. By translating minimal numerical values into perceptible visual contrast, the log transformation underscores emergent patterns of morphological expansion or stagnation in a statistically robust way.

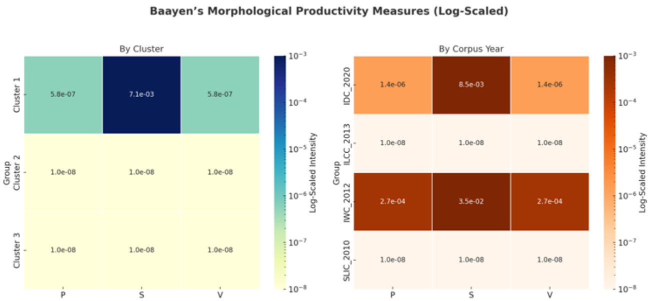


Figure 9. Baayen's morphological productivity measure.

The heatmap visualization reveals stark contrast in morphological productivity across the three LVC clusters. Notably, Cluster 1 stands out as the only group exhibiting non-zero values across all three productivity dimensions, with $P \approx 5.79 \times 10^{-7}$, $S = 0.0071$, and $V \approx 5.79 \times 10^{-7}$. These values, while low in absolute terms, are significant within the domain of morphological analysis, especially when compared to Cluster 2 and Cluster 3, both of which show zero productivity on all dimensions. The expanding productivity (S) in Cluster 1 is especially noteworthy, indicating that approximately 0.71% of the cluster's LVC types occur only once—suggesting latent morphological creativity. This aligns with prior findings from Yule's K and K-means Clustering, which also identified Cluster 1 at the most lexically diverse and least formulaic grouping. In contrast, Clusters 2 and 3 are likely composed fixed, repetitive constructions with entrenched morphological forms that do not generate novel expressions. The sharp visual contrast in the heatmap (*see* Figure 9), with deep coloration for Cluster 1 and near absence of color for the others, makes this differentiation immediately salient.

The corpus-based heatmap further demonstrates how morphological productivity has varied over time. IWC_2012 emerges as the most morphologically innovative corpus segment, with $P \approx 2.74\times10^{-4}$, $S = 0.0350$, and $V \approx 2.74\times10^{-4}$. These values indicate that 3.5% of its LVC types are hapax legomena, a clear marker of linguistic expansion and novel lexical information. This peak in productivity may reflect shifts in register or discursive style in 2012—possible tied to digital communication, media, or educational texts that foster lexical experimentation. IDC_2020, though far less productive than IWC_2012, still shows non-zero values ($P \approx 1.45\times10^{-8}$, $S=0.0085$, $V \approx 1.45\times10^{-8}$), suggesting a re-emergence of morphological innovation in the most recent data. Meanwhile, SLIC_2010 and ILCC_2013 exhibit no morphological productivity, with all values at zero. This finding correlates with the higher lexical concentration identified in ILCC_2013 (Yule's $K \approx 65.01$), which points to formulaic or standardized usage patterns.

Lastly, an examination through Shannon Entropy. Shannon Entropy offers a foundational information-theoretic metric for quantifying lexical unpredictability and distributional evenness within language data.[4] The entropy analysis at the cluster levels reveals substantial variation in the distributional structure of LVC usage. Cluster 1, with a Shannon Entropy of 8.47, emerges as the most information-rich and lexically diverse group. This high entropy value suggests that LVCs within this cluster are used with a relatively balanced frequency. In contrast, Cluster 2 registers the lowest entropy score of 4.72, reflecting a heavy reliance on a small set of highly frequent LVCs—an indicator of linguistic formulaicity and predictability. Cluster 3, with an entropy value of 5.62, occupies an intermediate position, implying a mixture of concentrated and distributed usage patterns. These entropy levels align closely with findings from previous metrics: Cluster 1 was earlier shown to have the lowest Yule's K (≈40.08) and the only cluster with non-zero Baayen productivity values, all of which point to high lexical diversity and creativity. The entropy heatmap further highlights this distributional divergence, with Cluster 1 displaying deep coloration indicative of information density, while Clusters 2 and 3 appear paler, visually reinforcing their restricted lexical behavior (*see* Figure 10). This confirms that the structural segmentation of LVCs corresponds to functionally distinct regimes of lexical usage.

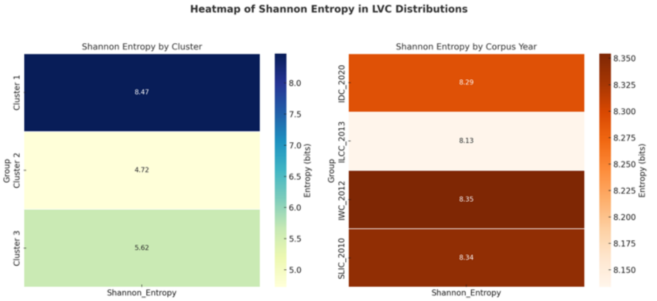


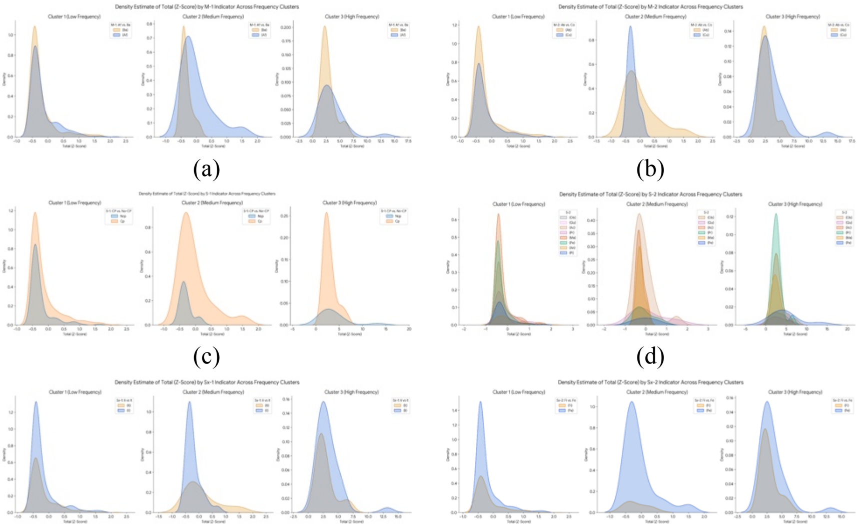
Figure 10. Heatmap of Shannon Entropy in LVC distribution.

When examined diachronically, entropy values across the four corpora exhibit smaller but still meaningful variation. The highest lexical unpredictability is found in IWC_2012 (H=8.35) and SLIC_2010 (H=8.34), closely followed by IDC_2020 (H=8.29). These high entropy values suggest that across these years, LVCs were used with considerable distribution balance, reflecting a lexicon characterized by frequent lexical alternation rather than fixed or formulaic usage. In contrast, ILCC_2013 exhibits the lowest entropy at 8.13, indicating a subtle but notable shift toward lexical concentration. This finding is consistent with ILCC_2013's previously observed high Yule's K (≈65.01) and zero Baayen productivity, further confirming its more repetitive and less generative linguistic structure. Although the numerical range between the highest and lowest entropy is relatively narrow (approximately 0.22), the log-scaled heatmap makes these differences visually salient, revealing how slight shifts in entropy can signal larger patterns of communicative regularity or innovation. These findings suggest that while lexical diversity remains relatively stable across time, certain corpus periods may experience micro-level contractions in expressive variability, possibly due to genre dominance or sociolinguistic standardization.

[4] For further readings, refer to Arora *et al.* (2022), Bentz *et al.* (2017), Diessel (2017), Pilgrim *et al.* (2024) and Shannon (1948, 1951).

3.3 Distribution of Indonesian LVCs across predefined linguistic conditions

First, distribution of LVCs in relation to nominal morphological features. Regarding the distribution of LVCs across nominal morphological features, Figure 11 offers a kernel density estimate (KDE) plot.[5] The presented graphical representation illustrates the KDE of total scores' LVCs (z-score) across three distinct frequency clusters, categorized by the morphological attributes of Affixed (Af) and Base (Ba) forms. The visualization (*see* Figure 11(a)) approximates the probability density function of total scores' LVCs, thereby providing a smoothed depiction of distribution within each cluster. The x-axis represents the total score, while the y-axis denotes the density, reflecting the relative likelihood of observing specific scores values. The differentiation between 'Af' and 'Ba' categories is visually conveyed through shaded regions within each cluster's subplot. To substantiate these observed distributional differences, a Chi-Square test of independence was performed. The resultant significant Chi-Square statistic (χ^2 = 10.2946, df = 2, p = .0058) compellingly demonstrates a statistically significant association between M-1 Indicator and Frequency Cluster. This outcome necessitates the rejection of the null hypothesis, thereby providing robust statistical evidence that the distribution of 'Af' and 'Ba' form LVCs is not uniform across frequency clusters. Consequently, the preliminary visual observations are corroborated by quantitative statistical analysis, affirming the existence of distinct distributional divergences.



(a) (b)

(c) (d)

(e) (f)

Figure 11. Distribution of LVCs across six predefined linguistic conditions.

[5] Within the purview of distributional analysis of Indonesian LVCs across predefined linguistic conditions, Kernel Density Estimate (KDE) plots serve as a salient non-parametric method, statistical inference procedures applicable to data exhibiting non-normal distributional properties, for visualizing the underlying probability density function of continuous variables, such as total scores or frequency metrics (Gramacki, 2018; Silverman, 2018; Wasserman, 2004, 2006; Węglarczyk, 2018).

Second, distribution across conceptual skeleton of SUBSTANCE. Specifically, the analysis focuses on the conceptual skeleton of SUBSTANCE of these nouns, distinguishing between concrete (Co) and abstract (Ab) entities. Collectively, the density plots (*see* Figure 11(b)) illustrate a clear modulation of the total count distribution as a function of frequency cluster. The observed shifts in skewness, kurtosis, and range underscore the heterogenous nature of the data and suggest that frequency plays a significant role in shaping the distribution of the total count variable. To substantiate the observed distributional differences in the total count variable across frequency clusters and M-2 indicator categories, a Chi-Square test of independence was conducted. The analysis yielded a statistically significant association between the M-2 Indicator and Frequency Cluster variables ($\chi^2 = 8.291$, df = 2, p = .016). This result compels the rejection of the null hypothesis, thereby providing statistical evidence that the distribution of 'Ab' and 'Co' categories is not uniform across the frequency clusters. In other words, the proportions of 'Ab' and 'Co' vary significantly depending on whether the data falls into the low, medium, or high-frequency cluster. This quantitative finding supports the preliminary visual observations derived from the kernel density estimates, affirming the existence of distinct distributional divergences as a function of both the M-2 indicator and frequency.

Third, distribution across scale of synonymity. By examining the presence of counterpart (Cp) or absence of counterpart (Ncp) as the synonymous pairs for LVCs, this part of analysis uncovers the semantic constraints. Collectively, the kernel density estimates (*see* Figure 11(c)), coupled with the descriptive statistics, reveal a systematic modulation of the LVCs' total count distribution as a function of frequency cluster and S-1 indicator category. The observed variations in skewness, spread, and modal tendencies, along with the differences in mean and standard deviation across clusters, highlight the complex interplay between frequency, the S-1 indicator, and the magnitude of total counts. To substantiate these observed distributional differences, a Chi-Square test of independence was performed. The resultant significant Chi-Square statistic ($\chi^2 = 10.251$, df = 2, p = .006) compellingly demonstrates a statistically significant association between the S-1 Indicator and Frequency Cluster. This outcome necessitates the rejection of the null hypothesis, thereby providing statistical evidence that the distribution of Cp and Ncp categories is not uniform across frequency clusters.

Fourth, distribution across prototypicality. In this context, nouns are categorized into seven semantic classes: People (Pe), Plants (Pl), Animals (An), Material (Ma), Objects (Ob), Qualities (Qu), Action (Ac), and Processes (Pr). By examining the distribution of LVCs across these prototypicality categories, this part of analysis aims to identify the semantic constraints. Overall, the kernel density estimates (*see* Figure 11(d)) highlight a systematic modulation of the total count distribution as a function of frequency cluster and S-2 indicator category. The observed variations in skewness, spread, and modal tendencies, along with the differences in mean and standard deviation across clusters and categories, underscore the complex interplay between frequency, the semantic categories represented by the S-2 indicator, and the magnitude of total counts. To substantiate the distributional differences across frequency clusters, an ANOVA was conducted. The results revealed a statistically significant effect of Frequency Cluster on LVCs' Total Count (F = 721.05, p < .001). Furthermore, a Kruskal-Wallis test revealed a significant difference in the distribution of total count values across the S-2 categories (H = 35.60, p < .001), suggesting that the semantic nature of the indicator also plays role in shaping the observed distributions.

Fifth, distribution across transitivity. As illustrated in Figure 11(e), the kernel density estimates illustrate the distribution of LVCs' total-count values across three frequency clusters, categorized by the Sx-1 indicator, which differences between *inherently intransitive*

(‘Ii’) and *inherently transitive* (‘It’) verb constructions. The observed variations in skewness, spread, and modal tendencies, along with the differences in mean and standard deviation across clusters, underscore the interplay between frequency clusters, the syntactic categories represented by the Sx-1 indicator, and the magnitude of total counts. To further substantiate these observations, a Chi-Square test of independence was performed to assess the relationship between frequency cluster and Sx-1 indicator. The analysis yielded a non-significant result ($\chi^2 = 0.436$, df = 2, p = .804), indicating that the distribution of ‘Ii’ and ‘It’ categories is not significantly different across the frequency clusters. This suggest that while the *magnitude* and *distributional shape* of total counts vary across clusters, the *proportions* of inherently intransitive and inherently transitive constructions are relatively consistent across these frequency bands.

Lastly, distribution across valency frame. As illustrated in Figure 11(f), the kernel density estimates visualize the distribution of total count values across three frequency clusters, categorized by the Sx-2 indicator, which distinguished between fixed valency (‘Fi’) and flexible valency (‘Fe’) LVCs. The kernel density estimates reveal a systematic influence of frequency cluster on the distribution of LVCs’ total count values across fixed and flexible valency constructions. While ‘Fi’ constructions consistently exhibit higher average total counts, the distributional shapes and the degree of variability change considerably across the frequency spectrum. The high-frequency cluster is distinguished by substantially larger total count values and a broader spread of the data. To further validate these observations, a Chi-Square test of independence was performed to assess the relationship between frequency cluster and Sx-2 indicator. The test yielded a statistically significant result ($\chi^2 = 9.339$, df = 2, p = .009), indicating that the distribution of the fixed and flexible valency constructions is not uniform across the frequency clusters.

3.4 Verb and noun elements within Indonesian LVcs

First, verb element within Indonesian LVCs. In the context of Indonesian LVCs, the verb element, in conjunction with a nominal constituent, engenders semantically cohesive phrasal units. This chapter undertakes a detailed analysis of the verbal component within such constructions, identifying two principal tendencies that characterize the light verb category in Indonesian, as visually represented in Figure 12, which illustrates the distribution between what shall be termed *true light verbs* (TLVs) and *vague action verbs* (VAVs).

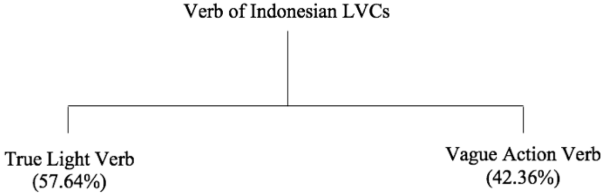


Figure 12. Hierarchical classification of verb elements in Indonesian LVC.

Notably, NB classifiers outperform other models in predicting optimal type-pairings, while Random Forest regression reveals that semantic parameters (S-1 and S-2) and corpus frequency patterns (especially in ILCC and IDC) play pivotal roles in verb classification. This analysis demonstrates that noun-types—particularly definite process and physiological state nouns—strongly predict the verb type. This approach affirms that verb behavior in LVCs is shaped by an interplay of aspectual neutrality, combinatorial selectivity, corpus distribution, and grammatical context.

Second, noun element within Indonesian LVCs. A pivotal component of analyzing Indonesian LVCs resides in the examination of their noun elements. The semantic properties of the noun heads within LVCs significantly influence the construction's overall meaning and usage patterns. The frequency and distribution analysis also reveal that there is a need for a detailed analysis of the noun element in those constructions. According to the current analysis of the aktionsart, Indonesian LVCs at least indicates the existence of two primary categories of noun element, as illustrated in Figure 13, namely *stative nouns* and *eventive nouns*. The stative nouns can be further classified into *psychological states* and *physiological states*, while eventive nouns can be classified into *indefinite process nouns*, *definite process nouns*, and *punctual nouns*.

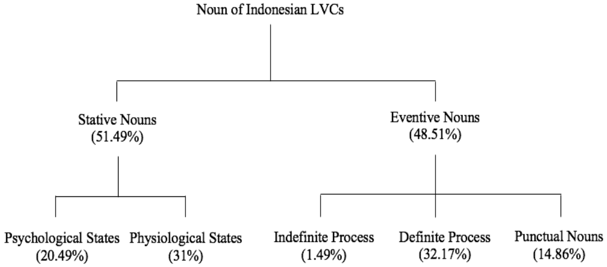

Noun of Indonesian LVCs
Stative Nouns (51.49%)
Eventive Nouns (48.51%)
Psychological States (20.49%)
Physiological States (31%)
Indefinite Process (1.49%)
Definite Process (32.17%)
Punctual Nouns (14.86%)

Figure 13. Hierarchical classification of noun elements in Indonesian LVC.

Through kernel density estimation, the chapter reveals frequency-based distinctions and divergence patterns across noun types. Importantly, machine learning models—particularly Random Forest—are employed to assess the predictive power of morphosyntactic and distributional features. Results show that eventiveness is strongly influenced by verb type and corpus distribution, while stative interpretations are more noun-driven. This chapter advances our understanding of how semantic weight, aspectual profile, and grammatical structure interact to shape LVC behavior in Indonesian.

4. Discussion

4.1. A short discussion of the three central hypotheses

The central hypotheses (*see* §2) posited in this study find substantial support in the empirical evidence gleaned from the corpus-based analysis of Indonesian LVCs. First, the examination into the cross-corpus frequency of linguistically defined LVCs, as detailed in the preceding section, yields several key insights relevant to Hypothesis 1. This hypothesis posited that, contrary to initial expectations of significant divergence due to inherent compositional disparities between the hypothetical and genuine LVC datasets, specific linguistically well-defined LVC types would exhibit statistically significant correlations in frequency across the four corpora under investigation, when controlled for genre and time period. The results of the correlation analyses provide support for this hypothesis. The observed correlations in the frequency of certain LVC type across the ILCC, SLIC, IDC, and IWC corpora suggest that, despite the differences in corpus composition and the temporal variations, there are underlying patterns of LVC usage that transcend corpus-specific idiosyncrasies. Specifically, the analysis of 'true light verb constructions', 'vague action verb constructions', 'eventive light verb constructions', and 'stative light verb constructions' reveals that some of these categories demonstrate a degree of consistency in their occurrence patterns.

Second, the investigation into the frequency distribution of LVCs as multiword expressions addresses Hypothesis 2, which predicts a deviation from a strict Zipfian distribution, particularly in the lower frequency range. The analysis, employing a combination of visual inspection (log-log plot), regression analysis, and goodness-of-fit tests, provides evidence supporting this hypothesis. While Zipf's Law, which posits an inverse relationship between word's frequency and its rank, is observed in many linguistic contexts, the behavior of LVCs presents a notable departure. The log-log plot analysis, a primary tool for visualizing Zipfian distributions, reveals that LVC frequency distribution does not conform to the expected linear pattern with a slope of -1. Instead, the plot demonstrates a curve, indicating a higher-than-predicted frequency of lower-ranking LVCs.

Third, the distributional analysis of LVCs across six predefined analytical dimensions directly addresses Hypothesis 3. This hypothesis posits that naturally occurring distinctive cluster of Indonesian LVCs exhibit statistically significant divergence in their distributional patterns across these dimensions, which are designed to capture variations at the morphological, semantic, and syntactic levels. The findings of this analysis, particularly when considered in the light of the three LVC clusters (Q1: Low Frequency, Q2: Medium Frequency, and Q3: High Frequency) derived from K-means clustering, and the KDE plots illustrating their distribution across the analytical dimensions, provide a nuanced perspective of LVC behavior. This stratification facilitates a more nuanced exploration of the factors influencing LVC distribution, enabling comparisons and contrasts across different levels of frequency, underscoring the crucial role these factors play in meaning construction (cf. Caro & Arús-Hita, 2020; Mattissen, 2023; Pompei, 2023). The K-means clustering effectively grouped LVCs based on their frequency characteristics, revealing inherent differences in how these constructions are employed within the language. This clustering, combined with the examination of the six analytical dimensions and the KDE plots, allows for a more granular understanding of LVC distribution. The KDE plots provide a visual representation of the probability density of LVCs across each dimension, highlighting both central tendencies and the spread of the data for each cluster.

4.2. On the asymmetrically bound structures of Indonesian LVCs

The results of conditional entropy analysis reveal a *marked asymmetry* in the internal dependency structure of LVCs (*see* Figure 14). The conditional entropy of the noun given the verb—H(Noun | Verb)—was calculated at 2.70 bits, indicating a relatively high degree of lexical variability: knowing the verb does not strongly predict which noun will follow. In contrast, the entropy of the verb given the noun—H(Verb | Noun)—was only 0.46 bits, suggesting a significant restriction in verb choice once the noun is specified. This implies that certain nouns co-occur consistently with particular verbs, forming *semantically entrenched templated* that exhibit strong selectional preference. These values suggest that, within Indonesian LVCs, nouns are the *collocational anchors,* while verbs exhibit greater contextual plasticity. This pattern aligns with the grammaticalization hypothesis, where light verbs lose semantic specificity and instead serve syntactic or aspectual functions, allowing a broad range or nominal complements. The visual representation through horizontal bar plot further accentuates this asymmetry, showing that the predictability of the verb slot is substantially higher than that of the noun slot. Such a pattern points to the entrenchment of certain noun–verb pairings, and the lexical creativity of verbs in accommodating semantic content.

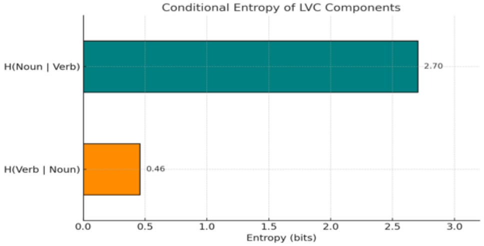


Figure 14. Conditional entropy of LVCs components.

This entropy asymmetry provides a quantitative foundation for characterizing the internal architecture of LVCs. The lower entropy in the verb slot given a noun suggest that many nouns are semantically tethered to a specific verb, reinforcing the notion that LVCs in Indonesian are often built around noun-governed collocational templates. This is consistent with prior linguistic theories that emphasize the unidirectionality of selectional constraints, wherein nouns evoke particular light verbs to realize specific eventive meanings (e.g., *'mengambil keputusan*' (to make a decision), '*memberikan bantuan*' (to give assistance)). From a morphosyntactic perspective, the result suggests that noun components serve as *semantic heads* in these constructions, while verbs increasingly function as *light lexical scaffolding*. The high entropy associated with nouns, meanwhile, implies that verbs are relatively unconstrained and can combine with a wide array of noun complements to accommodate stylistic, discursive, or genre-specific variation. This duality between *rigidity* and *flexibility* mirrors broader dynamics observed in multiword expressions and formulaic sequences across languages, where one constituent exhibits high collocational stability while the other contributes to variation.

5. Conclusion

5.1 Limitations of study

This study faces several data selection and corpus-genre diversity limitations. A narrow corpus may overlook spoken or informal language variations, while a larger, more varied corpus genre is needed to capture LVC distribution accurately. Challenges include the subjectivity in annotating LVCs due to the complex interaction between light verbs and their complements. Inconsistencies arise when distinguishing full verbs from LVCs, impacting the reliability of the findings. Despite the increased methodological complexities, a comprehensive approach involving diverse registers is essential for capturing the nuanced usage of LVCs across different contexts.

5.2 Recommendations

Future studies should broaden the scope of LVC analysis by employing a diachronic approach to understand their historical evolution, including grammaticalization and semantic changes. Deepening theoretical exploration through Cognitive Linguistics can clarify the mental processes behind LVC formation and interpretation, while Construction Grammar can analyze LVCs as form-meaning pairings with specific functions. Lastly, advanced corpus-analysis techniques, including pattern matching and idiomatic expression automatic-identification, will deepen insights. Integrating Discourse Analysis with genre-balanced corpora will enhance understanding of LVCs in real-world communication, accounting for context and shared knowledge.

References

Aikhenvald, A. Y. (2006). Classifiers and Noun Classes: Semantics. In *Encyclopedia of Language & Linguistics* (pp. 463–471). Elsevier. https://doi.org/10.1016/B0-08-044854-2/01111-1

Aikhenvald, A. Y. (2007). Typological distinctions in word-formation. In *Language Typology and Syntactic Description* (pp. 1–65). Cambridge University Press. https://doi.org/10.1017/CBO9780511618437.001

Aikhenvald, A. Y. (2017). A Typology of Noun Categorization Devices. In *The Cambridge Handbook of Linguistic Typology* (pp. 361–404). Cambridge University Press. https://doi.org/10.1017/9781316135716.012

Altmann, E. G., & Gerlach, M. (2016). Statistical Laws in Linguistics (pp. 7–26). https://doi.org/10.1007/978-3-319-24403-7_2

Altmann, G. (1985). Sprachtheorie und mathematische Modelle. *SAIS Arbeitsberichte Aus Dem Seminar Für Allgemeine Und Indogermanische Sprachwissenschaft* 8, 1–13.

Altmann, G. (1997). The Art of Quantitative Linguistics. *Journal of Quantitative Linguistics*, 4(1–3), 13–22. https://doi.org/10.1080/09296179708590074

Altmann, G. (2025). Language theory and mathematical models. In E. Kelih, J. Mačutek, & M. Koščová (Eds.), *Quantification in Linguistics and Text Analysis: Selected Papers of Gabriel Altmann*. De Gruyter Mouton.

Andreou, M. (2017). The semantics of compounding. *Morphology*, 27(4), 721–725. https://doi.org/10.1007/s11525-017-9311-1

Arora, A., Meister, C., & Cotterell, R. (2022). Estimating the Entropy of Linguistic Distributions. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 175–195. https://doi.org/10.18653/v1/2022.acl-short.20

Baayen, H. (1989). *A Corpus-Based Approach to Morphological Productivity. Statistical Analysis and Psycholinguistic Interpretation*. Vrije Universiteit.

Baayen, H. (1992). *Quantitative aspects of morphological productivity* (pp. 109–149). https://doi.org/10.1007/978-94-011-2516-1_8

Baayen, H. (2009). Corpus linguistics in morphology: Morphological productivity. In *Corpus Linguistics* (pp. 899–919). Mouton de Gruyter. https://doi.org/10.1515/9783110213881.2.899

Baayen, H., & Lieber, R. (1991). Productivity and English derivation: a corpus-based study. *Ling*, *29*(5), 801–844. https://doi.org/10.1515/ling.1991.29.5.801

Baker, P., & Egbert, J. (2016). Triangulating methodological approaches in corpus linguistic research. In *Triangulating Methodological Approaches in Corpus Linguistic Research*. https://doi.org/10.4324/9781315724812

Baker, P., Hardi, A., & McEnery, T. (2006). *Glossary of Corpus Linguistics*. Edinburgh University Press.

Bentz, C., Alikaniotis, D., Cysouw, M., & Ferrer-i-Cancho, R. (2017). The Entropy of Words—Learnability and Expressivity across More than 1000 Languages. *Entropy*, *19*(6), 275. https://doi.org/10.3390/e19060275

Biber, D. (2012). Corpus-Based and Corpus-driven Analyses of Language Variation and Use. In *The Oxford Handbook of Linguistic Analysis*. https://doi.org/10.1093/oxfordhb/9780199544004.013.0008

Binnick, R. I. (1991). *Time and the Verb*. Oxford University Press. https://doi.org/10.1093/oso/9780195062069.001.0001

Blevins, J. P. (2016). Word and Paradigm Morphology. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199593545.001.0001

Bygi, Z. R., Karimi-Doustan, G.-H., & Sharif, B. (2018). Explanation for alternating Light Verbs (LVs) in Persian complex predicates from a generative lexicon viewpoint. Language Related Research, 8(7), 429–452.

Caro, E. M., & Arús-Hita, J. (2020). Give as a light verb. *Functions of Language*, *27*(3), 280–306. https://doi.org/10.1075/fol.16036.mar

Cieri, C., Maxwell, M., Strassel, S. and Tracey, J. (2016). Selection criteria for low resource language programs. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4543–4549.

Coleman, L., & Kay, P. (1981). Prototype Semantics: The English Word Lie. *Language, 57*(1), 26. https://doi.org/10.2307/414285

Cruse, D. A. (1986). *Lexical Semantics*. Cambridge University Press.

Dal, G., & Namer, F. (2015). Frequency in morphology: For what usages? | La fréquence en morphologie: Pour quels usages? *Langages, 197*(1), 47–68. https://doi.org/10.3917/lang.197.0047

Diessel, H. (2017). Usage-Based Linguistics. In *Oxford Research Encyclopedia of Linguistics*. Oxford University Press. https://doi.org/10.1093/acrefore/9780199384655.013.363

Egbert, J., Larsson, T., & Biber, D. (2020). Doing linguistics with a corpus: Methodological considerations for the everyday user. In *Doing Linguistics with a Corpus: Methodological Considerations for the Everyday User*. https://doi.org/10.1017/9781108888790

Giparaitė, J. (2015). A Corpus-based Analysis of the Constructions Have/Take/Get a Bath and Have/Take/Get a Rest in British English. *Žmogus Ir Žodis*, *17*(3), 37–53. https://doi.org/10.15823/zz.2015.10

Giparaitė, J. (2024). A corpus-based analysis of light verb constructions with MAKE and DO in British English. *Kalbotyra*, *76*, 18–41. https://doi.org/10.15388/Kalbotyra.2023.76.2

Gramacki, A. (2018). *Nonparametric Kernel Density Estimation and Its Computational Aspects* (Vol. 37). Springer International Publishing. https://doi.org/10.1007/978-3-319-71688-6

Hopper, P. J., & Thompson, S. A. (1980). Transitivity in Grammar and Discourse. *Language, 56*(2), 251–299. https://doi.org/10.1353/lan.1980.0017

Hrenek, É. (2021). Synonymy in light verb constructions of the type feledésbe + verb | Szinonímia a feledésbe + ige típusú funkcióigés szerkezetek körében. *Magyar Nyelvor, 145*(3), 277–311. https://doi.org/10.38143/NYR.2021.3.277

Kettnerová, V. (2023). Valency structure of complex predicates with Light Verbs. In *Light Verb Constructions as Complex Verbs* (pp. 19–44). De Gruyter. https://doi.org/10.1515/9783110747997-002

Kittilä, S. (2006). The anomaly of the verb "give" explained by its high (formal and semantic) transitivity. *Linguistics*, *44*(3), 569–612. https://doi.org/10.1515/LING.2006.019

Köhler, R., Altmann, G., & Piotrowski, R. (2005). *Quantitative Linguistik*. Walter de Gruyter. https://doi.org/10.1515/9783110155785

Lieber, R. (2004). *Morphology and Lexical Semantics*. Cambridge University Press.

Lieber, R. (2010). *On the lexical semantics of compounds* (pp. 127–144). https://doi.org/10.1075/cilt.311.11lie

Lieber, R., & Štekauer, P. (2011). *Introduction: Status and Definition of Compounding*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199695720.013.0001

Magueresse, A., Carles, V., and Heetderks, E. (2020). Low-resource Languages: A Review of Past Work and Future Challenges. *Computation and Language*. https://doi.org/10.48550/arXiv.2006.07264

Malchukov, A. L. (2006). Transitivity parameters and transitivity alternations. In *Case, Valency and Transitivity* (pp. 329–357). https://doi.org/10.1075/slcs.77.21mal

Manin, D. Y. (2009). Mandelbrot's Model for Zipf's Law: Can Mandelbrot's Model Explain Zipf's Law for Language? *Journal of Quantitative Linguistics, 16*(3), 274–285. https://doi.org/10.1080/09296170902850358

Mattissen, J. (2023). Light Verbs and 'light nouns' in polysynthetic languages. In *Light Verb Constructions as Complex Verbs* (pp. 275–304). De Gruyter. https://doi.org/10.1515/9783110747997-011

Maxwelll-Smith, Z., Kohler, M., & Suominen, H. (2022). Scoping natural language processing in Indonesian and Malay for education applications. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, 171–228. https://doi.org/10.18653/v1/2022.acl-srw.15

McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge University Press. https://doi.org/10.1017/CBO9780511981395

Mel'čuk, I. (2006). *Aspects of the Theory of Morphology*. Mouton de Gruyter.

Mel'čuk, I. (2022). Support (=Light) Verbs. *Neophilologica, 34*, 1–30. https://doi.org/10.31261/NEO.2022.34.03

Miyamoto, T., & Kishimoto, H. (2016). Light verb constructions with verbal nouns. In *Handbook of Japanese Lexicon and Word Formation* (pp. 425–458). De Gruyter. https://doi.org/10.1515/9781614512097-016

Nagy T., István, Rácz, A., & Vincze, V. (2020). Detecting light verb constructions across languages. *Natural Language Engineering*, *26*(3), 319–348. https://doi.org/10.1017/S1351324919000330

Nugraha, D. S. (2024). A Corpus-Based Study of Memberi 'Give' Light Verb Constructions. *International Journal of Society, Culture and Language, 12*(2), 104–120. https://doi.org/10.22034/ijscl.2024.2023112.3389

Nugraha, D. S., & Vincze, V. (2024). Towards an Empirical Understanding of membawa 'bring': Corpus Insights into Indonesian Light Verb Constructions. *Jurnal Arbitrer*, 11(3), 278–296. https://doi.org/10.25077/ar.11.3.278-296.2024

Oakes, M. (2019). *Statistics for Corpus Linguistics*. Edinburgh University Press. https://doi.org/10.1515/9781474471381

Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review, 21*(5), 1112–1130. https://doi.org/10.3758/s13423-014-0585-6

Pilgrim, C., Guo, W., & Hills, T. T. (2024). The rising entropy of English in the attention economy. *Communications Psychology*, *2*(1), 70. https://doi.org/10.1038/s44271-024-00117-1

Pompei, A. (2023). How light is "give" as a Light Verb? A case study on the actionality of Latin Light Verb Constructions (with some references to Romance languages). In *Light Verb Constructions as Complex Verbs: Features, Typology and Function*. https://doi.org/10.1515/9783110747997-006

Ronan, P. (2014). Light verb constructions in the history of English. In *Studies in Corpus Linguistics* (Vol. 63). https://doi.org/10.1075/scl.63.05ron

Ronan, P. (2019). Simple versus Light Verb Constructions in Late Modern Irish English Correspondence: A Qualitative and Quantitative Analysis. *Studia Neophilologica*, *91*(1), 31–48. https://doi.org/10.1080/00393274.2019.1578182

Ronan, P., & Schneider, G. (2015). Determining light verb constructions in contemporary British and Irish English. *International Journal of Corpus Linguistics, 20*(3), 326–354. https://doi.org/10.1075/ijcl.20.3.03ron

Sebastian, D., Purnomo, H. D., & Sembiring, I. (2022). BERT for Natural Language Processing in Bahasa Indonesia. *2022 2nd International Conference on Intelligent Cybernetics Technology & Applications (ICICyTA)*, 204–209. https://doi.org/10.1109/ICICyTA57421.2022.10038230

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Shannon, C. E. (1951). Prediction and Entropy of Printed English. *Bell System Technical Journal*, *30*(1), 50–64. https://doi.org/10.1002/j.1538-7305.1951.tb01366.x

Silverman, B. W. (2018). *Density Estimation for Statistics and Data Analysis*. Routledge. https://doi.org/10.1201/9781315140919

Singh, A.K. (2008). Natural language processing for less privileged languages: Where do we come from? where are we going? In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*.

Singleton, D. (2016). *Language and the Lexicon.* Routledge. https://doi.org/10.4324/9781315824796

Tsvetkov, Y. (2017). *Opportunities and challenges in working with low-resource languages*. Carnegie Mellon University.

Vincze, V. (2011). *Semi-Compositional Noun + Verb Constructions: Theoretical Questions and Computational Linguistic Analyses*. University of Szeged.

Vincze, V. (2014). Valency Frames in a Hungarian Corpus. *Journal of Quantitative Linguistics, 21*(2), 153–176. https://doi.org/10.1080/09296174.2014.882188

Wasserman, L. (2004). *All of Statistics.* Springer New York. https://doi.org/10.1007/978-0-387-21736-9

Wasserman, L. (2006). *All of Nonparametric Statistics.* Springer New York. https://doi.org/10.1007/0-387-30623-4

Węglarczyk, S. (2018). Kernel density estimation and its application. ITM Web of Conferences, 23, 00037. https://doi.org/10.1051/itmconf/20182300037

Xu, H., Jiang, M., Lin, J., & Huang, C.-R. (2022). Light verb variations and varieties of Mandarin Chinese: Comparable corpus driven approaches to grammatical variations. *Corpus Linguistics and Linguistic Theory*, *18*(1), 145–173. https://doi.org/10.1515/cllt-2019-0049

Zipf, G. K. (2013). *The Psycho-Biology of Language*. Routledge. https://doi.org/10.4324/9781315009421

Appendices

A. Publications pertaining to the topic of the dissertation

Nugraha, D. S. (2024). A Morphological Analysis of the Indonesian Suffixation: A Look at the Different Types of Affixes and Their Semantic Changes. *GEMA, 24*(4), 109–132. http://doi.org/10.17576/gema-2024-2404-07 [D1|Q1 Scopus]

Nugraha, D. S. (2024). A Corpus-Based Study of *Memberi* "Give" Light Verb Constructions. *International Journal of Society, Culture & Language, 12*(2), 104–120. http://doi.org/10.22034/ijscl.2024.2023112.3389 [Q1 Scopus]

Nugraha, D. S. (2024). Analyzing Prefix /me(N)-/ in the Indonesian Affixation: A Corpus-Based Morphology. *Theory and Practice in Language Studies, 14*(6), 1697–1711. http://doi.org/10.17507/tpls.1406.10 [Q2 Scopus]

Nugraha, D. S., & Vincze, V. (2024). Towards an Empirical Understanding of *Membawa* "bring": Corpus Insights into Indonesian Light Verb Constructions. *Jurnal Arbitrer, 11*(3), 278–296. http://doi.org/10.25077/ar.11.3.278-296.2024 [No-Q Scopus]

Nugraha, D. S. (2024). Morphosemantic Features of *Memenuhi* "meet" in the Light Verb Constructions of Indonesian. *Linguistik Indonesia, 42*(2), 461–480. http://doi.org/10.26499/li.v42i2.638

Nugraha, D. S. (2023). Morphosemantic Features of *Membuat* "make" in the Light Verb Constructions of Indonesian. *LiNGUA, 17*(2), 131–142. http://doi.org/10.18860/ling.v17i2.17757

Nugraha, D. S. (2023). Morphosemantic Features of *Mengambil* "take" in the Light Verb Constructions of Indonesian. *International Journal of Linguistics and Translation Studies, 4*(3), 120–138. http://doi.org/10.36892/ijlts.v4i3.327

Nugraha, D. S. (2023). Morphosyntactic Features of *Membuat* "make" in the Light Verb Constructions of Indonesian. *European Journal of Language and Culture Studies, 2*(2), 33–43. http://doi.org/10.24018/ejlang.2023.2.2.80

Nugraha, D. S. (2022). Identifying Light Verb Constructions in Indonesian: A Direct Translation Approach. *International Journal of Language and Literary Studies, 4*(3), 298–311. http://doi.org/10.36892/ijlls.v4i3.1042

Nugraha, D. S. (2024). Some Notes on Indonesian Word Formation: A Study Based on the Derivational Morphology Approach. *South Asian Research Journal of Humanities and Social Sciences, 6*(1), 20–31. http://doi.org/10.36346/sarjhss.2024.v06i01.004

Nugraha, D. S. (2024). Investigating the Unproductive Morphological Forms in Indonesian Language. *Asian Journal of Education and Social Studies, 50*(4), 280–294. http://doi.org/10.9734/ajess/2024/v50i41330

B. Other Publications

Nugraha, D. S. (2024). Navigating Challenges and Opportunities: Incorporating Multimodal Analysis into Corpus Linguistics for Social Media Research. In *Corpora for Language Learning: Bridging the Research-Practice Divide.* Routledge. (pp. 36–38). https://doi.org/10.4324/9781003413301 [Scopus]

Nugraha, D. S. (2024). A Tale of Two Presidents: Indonesian Humor as Depicted in Political Cartoons. In *Communicating Political Humor in the Media.* Springer (pp. 45–71). http://doi.org/10.1007/978-981-97-0726-3_3 [Scopus]

Nugraha, D. S. (2022). Incorporating Cross-Cultural Competence into the ISOL Programme through Cultural-Based Materials and Corpus-Based Approach. *European Journal of Education and Pedagogy, 3*(3), 91–96. http://doi.org/10.24018/ejedu.2022.3.3.350